Molecular BioSystems

RSC Publishing

PAPER

Finding the targets of a drug by integration of gene expression data with a protein interaction network†

Cite this: *Mol. BioSyst.*, 2013, **9**, 1676

Griet Laenen,[ab] Lieven Thorrez,[ac] Daniela Börnigen[d] and Yves Moreau*[ab]

Received 12th October 2012,
Accepted 8th February 2013

DOI: 10.1039/c3mb25438k

www.rsc.org/molecularbiosystems

Polypharmacology, which focuses on designing drugs that bind efficiently to multiple targets, has emerged as a new strategic trend in today's drug discovery research. Many successful drugs achieve their effects via multi-target interactions. However, these targets are largely unknown for both marketed drugs and drugs in development. A better knowledge of a drug's mode of action could be of substantial value to future drug development, in particular for side effect prediction and drug repositioning. We propose a network-based computational method for drug target prediction, applicable on a genome-wide scale. Our approach relies on the analysis of gene expression following drug treatment in the context of a functional protein association network. By diffusing differential expression signals to neighboring or correlated nodes in the network, genes are prioritized as potential targets based on the transcriptional response of functionally related genes. Different diffusion strategies were evaluated on 235 publicly available gene expression datasets for treatment with bioactive molecules having a known target. AUC values of up to more than 90% demonstrate the effectiveness of our approach and indicate the predictive power of integrating experimental gene expression data with prior knowledge from protein association networks.

Introduction

For many years, the development of highly selective compounds interacting with a single target has been a pragmatic criterion in drug design. However, this reductionist approach is now being questioned, as its application has been concurrent with an increased rate of drugs failing in the late clinical development stages. Moreover, several lines of evidence suggest that many effective drugs act through the modulation of multiple targets and that this 'polypharmacology' can even be a therapeutic requirement.[1] While drug interaction with a set of targets seems to be beneficial for efficacy in many cases, there are also proteins one would rather avoid modulating,[2] for unintended drug activity at off-targets is one of the leading causes of unexpected side effects.[3] Off-target toxicity accounts for about one third of the failures in clinical development[1] and even causes withdrawal of approved drugs. Fenfluramine, for instance, was used in combination with phentermine as the popular appetite suppressant Fen-Phen, but was withdrawn from the market in 1997 because of causing valvular heart disease, later attributed to off-target activation of the serotonin receptor 2B (HTR2B).[4] A better knowledge of drugs' off-target interactions could prevent such serious adverse drug reactions and allow safer drug candidates to be prioritized for preclinical development.[5] Although usually unwanted and harmful, occasionally off-target effects can be beneficial and lead to new therapeutic indications. For example mifepristone, a progesterone receptor antagonist originally developed as an abortifacient, was repositioned as a treatment for hyperglycemia in adults with Cushing's Syndrome on the basis of its interaction with the glucocorticoid receptor.[6] Repurposing administration-approved drugs can not only speed up drug discovery and reduce costs through the elimination of toxicological and pharmacokinetic assessments,[4] it also holds promise for rare and neglected diseases,[7] as illustrated by mifepristone. Therefore, target prediction of approved drugs is highly valued in drug discovery nowadays.

The experimental testing of drug–target interactions remains cost- and time-intensive, and thus there is a strong incentive to develop *in silico* methods addressing this issue.[8] Several computational approaches have been developed over the years[9] exploiting molecular features of chemicals and proteins[10–17] as well as phenotypic effects of drug treatment.[17–26]

[a] *KU Leuven, Department of Electrical Engineering-ESAT, SCD-SISTA, Leuven, Belgium. E-mail: yves.moreau@esat.kuleuven.be*
[b] *iMinds Future Health Department, Leuven, Belgium*
[c] *KU Leuven, Department of Development and Regeneration @ Kulak, Kortrijk, Belgium*
[d] *Harvard University, Harvard School of Public Health, Biostatistics Department, Boston, MA, USA*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c3mb25438k

This journal is © The Royal Society of Chemistry 2013

The traditional computational methods can be classified as either ligand-based or structure-based. Ligand-based approaches like QSAR (quantitative structure–activity relationship) and similarity search compare a compound to molecules with an experimentally determined biological activity or to the known ligands of a target protein, typically using machine learning techniques.[10–13] Structure-based or docking-based approaches on the other hand, rely on the three-dimensional structure of the target to determine its interactions with small molecules.[14–16] More recently, the completion of the Human Genome Project has led to the emergence of a plethora of genomic tools including chemogenomics, which combines chemistry with genomics in order to predict ligand–target interactions on a larger scale.[27] In this spirit Zhao and Li[17] developed the computational framework drugCIPHER for predicting genome-wide drug–target interactions by relating pharmacological space, in terms of chemical and/or therapeutic similarities, to the relevance of targets based on protein–protein interactions in genomic space. Phenotype-based systems biology approaches like this are becoming more prominent due to the growing availability of phenotypic information on drugs in public databases such as DrugBank,[28] PharmGKB,[29] PROMISCUOUS,[30] SIDER,[31] and the Connectivity Map (cMap).[18,19] Besides therapeutic similarities among drugs, side effect similarities can also be exploited in this context, as illustrated by the work of Campillos *et al.*[20] on target identification *via* similarities between side effects summarized in the package inserts of marketed drugs. Finally, the prediction of drug–target interactions has also been attempted based on gene expression profiles following drug treatment. This approach is particularly appealing because it does not require any prior information on the compounds of interest.[21] One of the most remarkable developments in this domain is the Connectivity Map,[18,19] a large collection of gene expression profiles derived from treatment of human cell lines with over 1000 different small molecule perturbagens. A simple pattern matching algorithm compares these expression profiles with the expression pattern for a specific mutation, disease, or treatment of interest in order to search for functional associations with compounds from the cMap. Inspired by this strategy, Iorio *et al.*[21,22] merged the cMap expression profiles from treatments with the same drug on different cell lines, whereupon they constructed a drug network and associated tool MANTRA to predict a drug's targets based on gene set enrichment analysis of the drug's gene signature. Likewise Hu and Agarwal[23] constructed a disease–drug network that can be used to gain insight into the mechanism of action of both drugs and diseases, to explain side effects of drugs, and to explore the possibilities for drug repositioning. The major limitation of such drug and disease–drug networks is that inference of mode of action and biological effects is limited to compounds and diseases present in the network. Nevertheless, the group of Butte has reported the successful repositioning of two FDA-approved drugs by comparing the gene expression signatures for marketed drugs from the cMap to signatures for diseases extracted from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO). Both the prediction of the antiulcer drug cimetidine as a candidate therapeutic in the treatment of lung adenocarcinoma[24] and of the anticonvulsant topiramate as a novel therapeutic option for inflammatory bowel diseases[25] were experimentally validated. Aside from comparing expression profiles, one can also predict drug–target interactions by the analysis of differential expression at the level of a molecular network. Cosgrove *et al.*[26] developed a sparse simultaneous equation model and associated inferential strategy for predicting directly perturbed genes from expression profiles. However, within this framework a directly targeted gene is interpreted as a gene that, adjusted for the regulatory effects of other genes, shows an elevated expression level. This interpretation led to an inconsistent performance for drug target prediction.

Gene expression data allow for a straightforward determination of which genes are active and so this type of data is often being used to hint at a drug's mode of effect. Differential expression between drug treated cells or tissues and non-treated controls pinpoints the main transcriptional effects of a drug. Treatment does, however, not necessarily result in a significant increase or decrease of the targets' mRNA level. Nevertheless we believe that genes functionally related to a target do tend to be up- or down-regulated after treatment, since they are often involved in the same pathways and may undergo regulatory effects from both the target and molecules functioning downstream. On the basis of this assumption we propose to predict a drug's targets by considering gene expression following treatment in the view of a functional protein association network. More specifically, we prioritize genes as potential targets based on the differential expression of functionally related genes in the surrounding network module. Nitsch *et al.*[32] have applied a similar approach to select the most promising genes from a list of candidate disease genes in human genetics. They evaluated three distinct random walk-based machine learning techniques that diffuse differential expression values over a protein interaction or functional association network in order to score and prioritize candidate genes based on their differentially expressed neighborhoods. The three strategies, being kernel ridge regression ranking, heat kernel diffusion ranking, and Arnoldi diffusion ranking, are all based on the exponential diffusion kernel, but differ in their exact methodologies. The best results were obtained by iterative heat kernel diffusion of RMA preprocessed expression data over the functional STRING network.[33] In our study we extended this kernel diffusion methodology and assessed its performance on a genome-wide drug target prioritization problem. Additionally, we also introduce a new diffusion method, based on connectivity correlations in the protein network. Drug target prioritization by means of both diffusion strategies was evaluated on a large benchmark consisting of 235 publicly available gene expression datasets for treatment with bioactive molecules that have a known target. For these data AUC values of up to 92% could be obtained.

Materials and methods

Functional protein association network

A functional protein association network can be modeled as an undirected graph $G = (V, E)$, where V is the set of vertices or

nodes and E is the set of edges. Two nodes, representing proteins, are connected by an undirected edge if there is a functional association between them. Each edge can be assigned a weight resembling the strength of evidence for the association.

Such a functional protein association network for human was obtained from STRING,[33] a database of known as well as predicted protein–protein associations, including both direct (physical) and indirect (functional) interactions. STRING integrates prior knowledge from OMIM[34] and PubMed abstracts[35] with associations from interaction databases such as MINT,[36] HPRD,[37] BIND,[38] DIP,[39] BioGRID,[40] IntAct,[41] KEGG,[42] Reactome,[43] NCI/Nature PID,[44] and Gene Ontology.[45] These known interactions are further complemented with newly predicted interactions based on gene fusion events, conserved genomic neighborhoods, phylogenetic co-occurrence, and co-expression. Lastly, interactions are also transferred between organisms by means of exhaustive cross-genome homology searches. As of version 9.0, these protein–protein similarity data are imported from the Similarity Matrix of Proteins (SIMAP) Project.[46] All transferred interactions as well as all predicted and imported associations are provided with a probabilistic confidence score, derived by benchmarking the associations against the functional grouping of proteins as annotated at KEGG. The associations are scored independently for the various major data sources, after which a combined score is computed indicating the higher confidence when more types of information support a given association.

We extracted a functional association network from two STRING releases: 8.3 and 9.0. The human protein network obtained from STRING 8.3 consists of 17 369 proteins and 1 288 886 interactions; the network extracted from STRING 9.0 consists of 18 600 proteins and 1 640 707 interactions. The protein Ensembl IDs from both these STRING networks were mapped to their encoding HGNC gene symbols.

Drug target prioritization strategies

Genome-wide knowledge on functional relations between proteins can help us achieve an understanding of the system as a whole. We evaluated two strategies integrating this information with gene expression data to establish a ranking of genes as potential drug targets and compared them to a simple expression ranking strategy, prioritizing genes based on expression data only. Both integrative methods assess the relevance of a gene as a drug target by considering the differential expression of its surrounding subnetwork of functionally related genes. The genes surrounding a drug target are expected to show an altered expression after drug administration in response to the perturbation of the target itself, while other genes, not functionally related to a target, are assumed not to be part of such a disrupted expression module. The subnetwork around each gene is selected either by means of a kernel-based random walk through the STRING network, or by means of computing connectivity correlations between node pairs.

1. **Simple expression ranking.** Ranking genes based on their own differential expression following treatment (Fig. 1A) can be a simple strategy to predict a drug's targets using only expression data. For each gene the expression level difference between the drug treated and untreated samples is assessed in terms of log2 ratios or moderated t-statistics, after which the genes can be ranked accordingly. This simple expression ranking can be considered a baseline for performance assessment of the

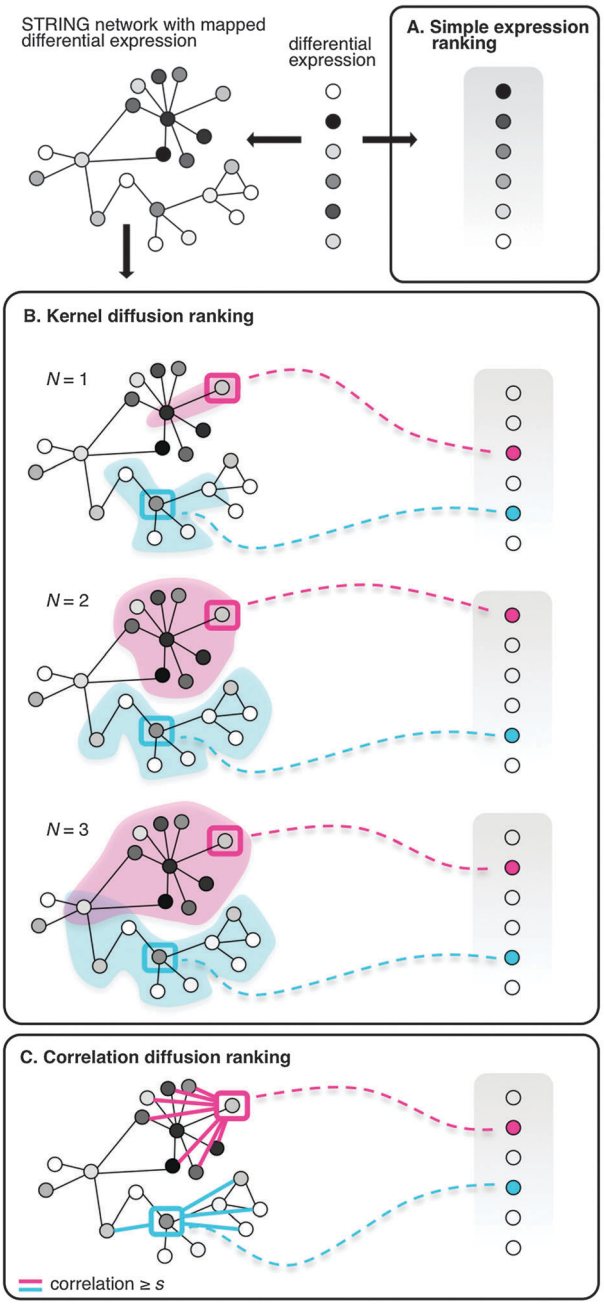


Fig. 1 The drug target prioritization strategies. A treatment *versus* control differential expression level is determined for every gene. (A) Based on these a simple expression ranking is set up as a baseline. To integrate knowledge on functional associations between proteins, the differential expression values are mapped to the STRING network. (B) Using kernel diffusion, genes are ranked based on the differential expression of their closest neighbors in this network. The number of steps through the network determines the size of the neighborhood taken into account. (C) The correlation diffusion method ranks the genes based on the differential expression of genes with a strong connectivity correlation.

This journal is © The Royal Society of Chemistry 2013

proposed network-based drug target prioritization strategies set out below.

2. Kernel diffusion ranking. For both the STRING 8.3 and STRING 9.0 network, represented as undirected graphs with symmetric weights w_{ij} between nodes i and j, the adjacency matrix A has entries $a_{ij} = w_{ij}$ if the nodes i and j are connected and $a_{ij} = 0$ otherwise. The combinatorial Laplacian matrix is given by $D - A$ where D is a diagonal matrix with $d_{ii} = \sum_j a_{ij}$. The asymmetric normalized version of the Laplacian is defined as $L_1 = I - D^{-1}A$, whereas the symmetric normalized version of the Laplacian, called the regularized Laplacian, is defined as $L_2 = I - D^{-1/2}AD^{-1/2}$.

Both these normalized Laplacians are used in a drug target prioritization method that diffuses the differential expression values of all genes over the STRING network based on the confidence scores of the neighboring functional associations. This diffusion process can be formulated using a Laplacian exponential diffusion kernel, defined by Kondor and Lafferty[47] as

$$K = e^{-\alpha L} = \lim_{n\to\infty}\left(I - \frac{\alpha}{n}L\right)^n,$$

which originates from the differential equation describing heat diffusion in the material.

To obtain a genome-wide drug target prioritization, a scoring vector p_α is computed by multiplying the differential expression values p_0 with this heat kernel:

$$p_\alpha = p_0\, e^{-\alpha L},$$

with L being one of the normalized Laplacian matrices. We have applied a discrete approximation of this heat kernel introduced by Yang *et al.*[48] and recently used by Nitsch *et al.*[32] to prioritize candidate disease genes:

$$p_\alpha = p_0\left(I - \frac{\alpha}{N}L\right)^N,$$

with N being the number of iterations and α the diffusion rate. By means of this iterative diffusion process the differential expression value for each gene is spread to all neighboring nodes reached in no more than N steps (Fig. 1B). All possible paths of length N or less are thereby taken into account, resulting in a faster diffusion between strongly interconnected nodes, which are assumed to have a more reliable functional relationship. This partial diffusion strategy can be considered comparable to the functional flow method proposed by Nabieva *et al.*[49] for protein function prediction. However, instead of applying a guilt-by-association principle requiring *a priori* knowledge, our method relies on experimental gene expression data. Furthermore, we introduce a diffusion parameter α controlling the diffusion rate. A value close to 1 causes rapid diffusion, giving a larger weight to genes located further away, while a value close to 0 promotes a slower diffusion process, giving a larger weight to nearby genes. As a result of this diffusion each node gets a new score, no longer just representing its own differential expression, but also including information on the differential expression of its close neighborhood. Drug targets are thus assumed to get a high score, since these genes tend to be the central nodes in subnetworks showing significant transcriptional changes following treatment.

In order to attenuate the systematic bias of the obtained scores toward highly connected genes, p-values indicating a gene's significance as a potential drug target are computed by means of a permutation procedure. The differential expression values are randomized across the network a 1000 times, after which a p-value is calculated for each gene as the fraction of permutations resulting in a score at least as extreme as the real score.

3. Correlation diffusion ranking. An alternative to the heat kernel approach, which diffuses differential expression values to neighboring nodes in the STRING network, is a correlation-based approach diffusing the differential expression values to genes/proteins considered functionally related based on the correlation of their weighted connectivity structure (Fig. 1C). To this end a Pearson correlation coefficient ρ is computed for each pair of nodes:

$$\rho_{ij} = \frac{\mathrm{cov}(w_i, w_j)}{\sigma_{w_i}\sigma_{w_j}},$$

with w_i the vector containing weights for the interactions between node i and all other nodes. By filtering out all connectivity correlation coefficients below a certain threshold s and normalizing the coefficients based on the remaining correlations per gene, the elements of a normalized correlation matrix C can be defined as

$$c_{ij} = \begin{cases} \dfrac{\rho_{ij}}{\sum_j \rho_{ij}}, & \text{for all } \rho_{ij} \geq s \\ 0, & \text{otherwise.} \end{cases}$$

Multiplication of the differential expression values with this matrix results in a ranking score for each gene:

$$p_\alpha = p_0\, C.$$

The obtained scores are again corrected for the number of nodes taken into account by means of the randomization procedure described above.

Test data

Our drug target prioritization strategies were evaluated on 235 publicly available gene expression datasets (accession numbers are listed in Table S1, ESI†) where treatment with a bioactive molecule was tested against a control on one of three different Affymetrix Human GeneChips: HG U133A, HG U133 Plus 2.0, or HuGene 1.0 ST. The raw CEL files were downloaded from GEO[50] and preprocessed with the RMA function of the R/Bioconductor affy[51] or oligo[52] package. After preprocessing, we computed three different measures of differential expression for each gene using the R/Bioconductor limma[53] package: the log2 ratio, the moderated t-statistic, and the significant log2 ratio. This latter statistic was obtained by setting the log2 ratio of all genes with a moderated t-statistic-derived p-value larger than 0.05 to zero.

The 235 datasets account for 130 unique treatment molecules (listed in Table S1, ESI†) that do not only include small molecules, but also antibodies, other proteins and a nucleic acid. For each of these molecules we identified the primary target (listed in Table S1, ESI†) from the literature and databases such as PubChem,[54] ChEMBL,[55] and BindingDB,[56] after which we retrieved its position in the corresponding genome-wide ranking lists resulting from application of our prioritization methods.

ROC analysis

The performance of the different drug target prioritization approaches was assessed in terms of the sensitivity and specificity with which the primary targets appear at the top of the ranking lists. In an ROC (Receiver Operating Characteristic) curve, the true positive rate (sensitivity) was plotted against the false positive rate (1 – specificity) by imposing all possible cut offs on the prioritization lists, each time dividing the genes into positive and negative ranked genes. True positives are the primary targets ranked above the cut off in the corresponding prioritization list, whereas false positives are all other genes ranked above. False negatives are the primary targets below the cut off and true negatives all other genes below.

As an evaluation measure, we computed the AUC (Area Under the ROC Curve) which is equivalent to the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance, according to Hanley and McNeil.[57] In our case this means the AUC would be 100% if all primary targets ranked in the first position, while it would be only 50% if the prioritization was no better than ranking the genes randomly.

Results and discussion

In this section, we evaluate the ability of our proposed prioritization strategies to predict a drug's targets by combining information on differential gene expression following treatment with knowledge on functional associations between proteins. The performance is assessed by means of a ROC analysis on genome-wide target prioritizations, obtained by these methods on 235 gene expression datasets for bioactive molecules having a known target. The corresponding AUC values are presented in Table 1. Apart from the prioritization strategies themselves, also the two integrated data sources, being the functional protein association network and the gene expression dataset, are investigated for their impact on the performance level.

Performance of the drug target prioritization methods

1. **Simple expression ranking.** Upon target binding, drug treatment can give rise to a plethora of complex cellular responses, including a positive or negative feedback loop adjusting the expression level of the respective target mRNA. Hence genes are ranked according to their own differential expression following treatment in order to establish a baseline for comparison with the suggested drug target prioritization strategies. Only genes present in the STRING network are taken up in these rankings, so as to make a direct comparison with the different prioritization methods possible. Table 1 shows that this simple strategy is not sufficient to predict a drug's targets: AUC values vary between 64% and 66%. This is consistent with the findings of Iskar *et al.*[58] concerning the quantification of drug-induced alterations of target mRNA levels. They extracted gene expression data for drug treatment from the cMap[18,19] for 466 drugs; only 8% of the targets appeared to be subject to regulations at the mRNA level. This low percentage may, however, in part be attributed to the relative short time (6 hours) between the addition of the drug and gene expression analysis in the cMap approach,[18] during which there is not sufficient time for effects *via* feedback loops. In the 235 datasets used for our analysis, time between drug addition and gene expression analysis is more variable and mostly longer. Still, the simple expression ranking results indicate that drug treatment mostly leads to relatively small differences in mRNA expression of the direct targets. Hence the need to consider the neighborhood in order to find a drug's targets based on gene expression data.

2. **Kernel diffusion ranking.** The kernel diffusion strategy does not restrict itself to the actual differential expression of genes to set up their ranking as potential targets. That is to say, it also takes the differential expression of close neighbors in the STRING network into account. To what extent the contribution of these neighbors to the ranking score decreases with an increasing distance in the network, is determined by the diffusion rate α (see Materials and methods). Table S2 (ESI†) shows the performance of kernel diffusion on the test data using different values for this diffusion parameter. Results show a decrease in performance with the decrease of α. Since drug targets are not always differentially expressed following treatment, as shown by the simple expression ranking results, the contribution of the gene for which a score is computed has to be kept small in comparison with the contribution of its neighboring nodes. This demands a rapid diffusion and so a higher diffusion parameter of 0.9 is required.

Besides the degree of diffusion, the edge weights between neighboring nodes also play an important role in the contribution of each gene. Weights are extracted from the STRING network, after which they are normalized by computing one of the Laplacians discussed in the Materials and methods section. The asymmetric normalized Laplacian L_1 ensures that the weights of all edges leaving a node, including the self loop, sum up to one. In this way the weight used for computing the contribution of a node j to the score of a neighboring node i is inversely proportional to the node degree of node j, defined as the sum of the weights of all adjacent nodes. Thus a lower confidence is given to highly connected nodes and edges get different weights for both directions. Using the symmetric normalized Laplacian L_2 edge weights are considered the same in both directions. The contribution of a node j to the score of a neighboring node i is now divided by the square root of the degree of node j multiplied by the degree of node i.

Table 1 depicts the performance level of the kernel diffusion strategy for both Laplacians. The performance has improved significantly in comparison to simple expression ranking, with

 This journal is © The Royal Society of Chemistry 2013

Table 1 AUC values for the different drug target prioritization strategies

				STRING 8.3	STRING 9.0	Randomized network
Simple expression ranking			Log2 ratio	0.659	0.657	0.659
			t-Statistic	0.646	0.644	0.646
Kernel diffusion ranking	L_1	$N = 1$	Log2 ratio	0.835	0.785	0.646
			t-Statistic	0.828	0.767	0.642
			Sign. log2 ratio	0.821	0.780	0.626
		$N = 2$	Log2 ratio	0.829	0.801	0.663
			t-Statistic	0.822	0.792	0.651
			Sign. log2 ratio	0.830	0.805	0.644
		$N = 3$	Log2 ratio	0.820	0.797	0.662
			t-Statistic	0.812	0.784	0.651
			Sign. log2 ratio	0.825	0.801	0.644
	L_2	$N = 1$	Log2 ratio	0.909	0.903	0.646
			t-Statistic	0.911	0.901	0.641
			Sign. log2 ratio	0.887	0.876	0.625
		$N = 2$	Log2 ratio	0.828	0.822	0.662
			t-Statistic	0.814	0.806	0.650
			Sign. Log2 ratio	0.846	0.840	0.642
		$N = 3$	Log2 ratio	0.812	0.805	0.663
			t-Statistic	0.795	0.792	0.651
			Sign. Log2 ratio	0.840	0.834	0.644
Correlation diffusion ranking			Log2 ratio	0.919	0.893	0.657
			t-Statistic	0.919	0.891	0.608
			Sign. Log2 ratio	0.918	0.896	0.645

AUC values ranging from 76% up to 91%. Results are comparable for both Laplacians when taking two or three steps in the network, but the symmetric normalized Laplacian clearly outperforms the asymmetric normalized Laplacian when limiting the scores to the differential expression values of direct neighbors ($N = 1$). The higher AUC values for such a one-step diffusion based on the symmetric normalized Laplacian may reflect some typical properties of the STRING network. First of all STRING is not limited to known, physical interactions, but also includes predicted, functional relations. This may substantially reduce the reliability of indirect links between second and third level neighbors. Moreover, STRING is a scale-free network, with most proteins participating in only a few associations and a few hubs interacting with thousands of proteins. These hubs make the network suffer from the small world effect, that is that most nodes can be reached from every other node through a few steps only. Most likely too many nodes are reached in two or three steps, causing dilution of the differential expression signals of direct neighbors. In addition, network hubs tend to be essential proteins regulated through complex mechanisms, making them less suitable to capture a transcriptional response focused around a single gene. Thus the confidence of such highly connected nodes needs to be corrected accordingly, as is done by normalization of the Laplacian. Results suggest downweighting by the square root of the node degree using the symmetric normalized Laplacian performs best.

3. **Correlation diffusion ranking.** Table S3 (ESI†) illustrates the results obtained by applying correlation diffusion ranking with different correlation thresholds s (see Materials and methods). The AUC values clearly show that a lower threshold of 0.1 is required to obtain adequate rankings, presumably because of the reduced correlation matrix sparsity.

Although the kernel diffusion method achieved the best performance when taking into account direct neighbors only, results obtained for the correlation diffusion method with a correlation threshold of 0.1 indicate that drug target prioritization can benefit from the inclusion of differential expression measures for nodes lying further away in the network. Indeed, Table 1 shows that the best AUC value (91.9%) was derived by the application of correlation diffusion, which includes the differential expression of all nodes with a correlated connectivity structure for computing ranking scores. Nodes with a high degree of correlation can be interpreted here as nodes sharing a large part of their direct neighbors. The good performance when diffusing the expression signals to such correlated nodes could be explained by proteins associated with the same set of proteins having greater likelihood of sharing functional characteristics. For example, Chua *et al.*[59] observed that a substantial number of proteins in the *Saccharomyces cerevisiae* GRID interaction network share functions with second level but not first level neighbors. This may explain the higher reliability of functional associations based on connectivity correlations in comparison to associations derived by taking short paths through the network, as demonstrated by the higher AUC values for correlation diffusion ranking in comparison with two- or three-step kernel diffusion.

Impact of the network

An important bottleneck for the performance of both diffusion methods is the quality and coverage of the functional network around the actual drug targets. We have chosen to extract our network from the STRING database, since it combines multiple heterogeneous data sources, consequently mitigating sparsity and boosting reliability. Furthermore, the STRING network has

already proven to be effective for similar prioritization problems as described by Nitsch *et al.*[32] Our drug target prioritization methods were evaluated on the current STRING 9.0 network as well as on its previous release STRING 8.3, in order to investigate the effect of small network differences on the effectiveness of the proposed methods. Table 1 shows a direct comparison between both STRING releases and also includes results obtained with a randomly permuted STRING network. The performance of this last network is not completely random; AUC values compare to those obtained by simple expression ranking. This could be explained by the fact that the edges are permuted but that the initial expression value on each node remains the same. Therefore, the informative contribution of the expression level of a gene itself remains preserved during prioritization, such that the score of each gene is determined by its own differential expression plus the noise caused by the differential expression of its randomly assigned neighbors. When comparing the two STRING releases, the updated version 9.0 results in slightly worse AUC values regardless of the applied method. This observation indicates that although coverage tends to be an important property, it may yield a larger portion of false positive associations, lowering the contribution of the truly associated genes. Fig. 2 illustrates ROC curves for both versions of the STRING network, derived from correlation diffusion rankings and symmetric as well as asymmetric normalized Laplacian kernel diffusion rankings based on log2 ratios. For all three methods compared, AUC was 1–5% better when the STRING 8.3 version was used.

Last, we considered the dependency of the results on different protein association evidence types integrated in the STRING network: co-expression, association in a curated database, experimental data, co-occurrence across genomes, co-mentioning in PubMed abstracts, occurrence in close neighborhood across genomes, and gene fusion. The two best performing ranking methods, one-step symmetric normalized Laplacian kernel diffusion and correlation diffusion, were applied to seven variants of the STRING 8.3 network, each comprising a single evidence type. The obtained results are depicted in Table 2. As the AUC values for each type of evidence are only based on those datasets for which the drug target is present in the relevant network variant, not all results are equally representative. Nevertheless, text mining clearly outperforms all other kinds of functional associations. This is partially due to the larger number of genes and interactions included in the text mining network compared to the often small or sparse networks representing the other evidence types. The gene fusion network consists for example of only 3300 genes connected through only 4017 interactions of which most have a very low weight. As a consequence only a minimal diffusion can take place, resulting in AUC values comparable to those obtained by simple expression ranking. Integrating the different association types resolves such sparsity issues and provides us with a more complete and hence robust network for our diffusion approach. Moreover, the higher AUC values for text mining can also be partially attributed to a performance assessment based on the prediction of drug targets of which many are well studied disease genes abundant in the scientific literature. Therefore, integration of all other evidence types is vital for our method to also perform well in a more prospective research setting.

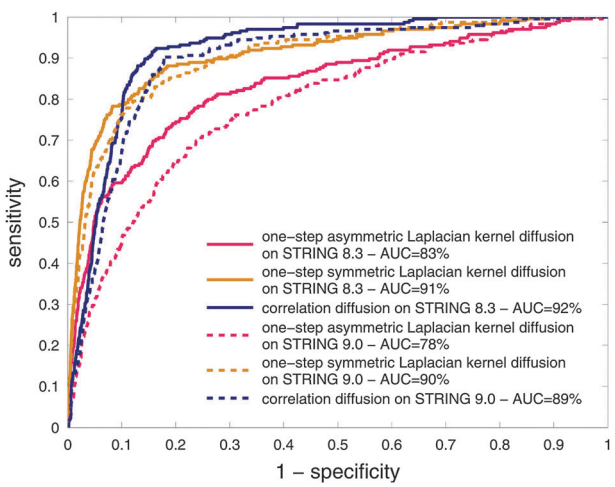


Fig. 2 ROC curves comparing the performance of the STRING 8.3 (solid line) and STRING 9.0 (dotted line) networks when the proposed prioritization methods are applied to log2 ratio statistics.

Impact of the expression data

Just like the functional network, the gene expression data, measured before and after treatment with a drug of interest, also have an important impact on the performance of the target prioritization methods. The quality of the measurements and the experimental set up (*e.g.* relevant cell type) can strongly influence the results. Hence divergent results were sometimes observed for the same drug tested in different experiments. Moreover, differential expression values can be passed to the methods in various manners. We compared rankings based on log2 ratios (for all genes as well as for the significant genes only) and moderated *t*-statistics. Table 1 shows that the performance differences between these expression measures are almost negligible. We also examined the influence of the type of experiment on the performance level, by splitting up the test data in *in vivo* and *in vitro* experiments. In the 27 *in vivo* experiments a patient group was treated with the drug of

Table 2 AUC values for the different protein association evidence types in STRING resulting from one-step symmetric normalized Laplacian kernel diffusion and correlation diffusion of log2 ratios

Protein association evidence type	Number of datasets tested	AUC for kernel diffusion ranking	AUC for correlation diffusion ranking
All	235	0.909	0.919
Co-expression	30	0.657	0.698
Database	230	0.768	0.773
Experimental	235	0.737	0.789
Co-occurrence	37	0.690	0.728
Text mining	235	0.906	0.920
Neighborhood	13	0.819	0.709
Fusion	111	0.605	0.673

This journal is © The Royal Society of Chemistry 2013

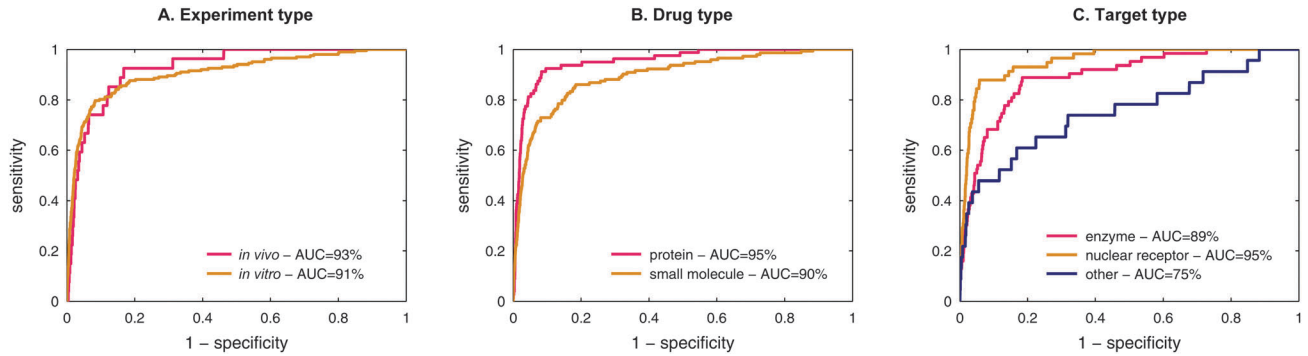


Fig. 3 ROC curves for a one-step symmetric normalized Laplacian kernel diffusion of log2 ratios from different data splits, illustrating the effect of the experiment, drug, and target type on the performance.

interest, in the 202 *in vitro* experiments the drug was administered to a cell line or biopsy, the 6 datasets derived from xenografts were not included in the analysis. Fig. 3A shows the ROC curves for a single-step symmetric normalized Laplacian kernel diffusion applied to the log2 ratios of both data subsets. With a difference of only 2% between both AUC values, the performance of our method seems quite robust for the type of experiment.

Comparison of the performance on different drug and target types

Together with the network characteristics and the quality of the expression data, also the prioritization problem in itself can affect the performance level of the diffusion methods. Specifically, the methods may not be equally effective for each type of drug and each type of target. Since most of our test datasets originate from treatment with either a protein (80, antibodies not included) or a small molecule (144), a simple partition of the two can elucidate whether the diffusion of differential expression measures over a functional protein network is more effective in predicting protein targets or small molecule targets. Fig. 3B shows that, with an AUC value of 95%, the one-step symmetric normalized Laplacian kernel diffusion performs slightly better for protein drugs, as compared to an AUC value of 90% for the small molecules. This may be explained by the fact that these proteins often are natural ligands of their targets, exhibiting both high specificity and high affinity.

Our prioritization methods are based on the assumption that drug treatment induces differential expression of genes that are functionally related to the actual drug target. In a last analysis we examine for which type of targets this assumption is most correct, by partitioning our small molecule test datasets according to the target class. We distinguish between enzymes (63), nuclear receptors (58), and others (23), the latter being a heterogeneous group consisting of, among others, membrane receptors, structural proteins, and secreted proteins. Fig. 3C shows that the performance of our one-step symmetric normalized Laplacian kernel diffusion is significantly lower for this last class in comparison to the other two classes. This may be because interference with the function of some of these proteins does not result in a direct, focused transcriptional change but in a delayed, more diffuse signal that is not picked up by our method. The nuclear receptors perform best with an AUC value of 95% and the enzyme class obtains an in-between AUC of 89%. As nuclear receptors are transcription factors causing gene expression changes directly downstream, it is not surprising that our method is most effective for this target type. By normalizing scores based on protein classification we might be able to correct for this target type bias.

Conclusion

To explain and prevent serious adverse drug reactions, to allow for safer drug candidates to be prioritized for preclinical development, and to explore the possibilities for drug repositioning, a better knowledge of drug–target interactions is vital. Several computational approaches tackling this problem have been described. We propose the use of gene expression data following treatment, despite the fact that direct targets are not necessarily retrieved by a standard differential expression analysis. Our approach considers the expression data in the context of a molecular network, thereby taking regulatory interactions between the proteins into account. More specifically, differential expression values following treatment are diffused over a functional subnetwork around each protein, resulting in a new statistic based on which a genome-wide ranking of potential targets can be established. We have explored two strategies for diffusion: a kernel-based random walk, employing a symmetric or asymmetric normalized Laplacian, and diffusion to genes with a strong connectivity structure correlation. Both methods were evaluated on 235 publicly available gene expression datasets for bioactive molecules with a known target. With AUC values of 91% and 92% respectively, the best results were obtained by single-step symmetric normalized Laplacian kernel diffusion and by correlation diffusion, both over the STRING 8.3 network. Results show no significant performance differences for the type of experiment from which gene expression data were derived, but performance does seem to depend on the type of drug, with protein drugs performing slightly better than small molecules. Our results also suggest performance differences between different target classes, depending on how concentrated the drug-induced transcriptional response

downstream of the target protein is. Still we believe that both proposed diffusion strategies show promise for future drug discovery research, in particular for drug target prioritization on a genome-wide scale.

Acknowledgements

YM and LT are professors at the Katholieke Universiteit Leuven, Belgium. GL is supported by a PhD grant of the Flemish Institute for Science and Technology (IWT). LT is a Postdoctoral Fellow of the Research Foundation – Flanders (FWO). Research is supported by Research Council KU Leuven: GOA MaNet, KUL PFV/10/016 *SymBioSys*, IOF/KP/12/010 Immunosuppression, Hercules Stichting: Hercules III PacBio RS, iMinds 2012.

References

1 A. L. Hopkins, *Nat. Chem. Biol.*, 2008, **4**, 682–690.
2 A. Koutsoukas, B. Simms, J. Kirchmair, P. J. Bond, A. V. Whitmore, S. Zimmer, M. P. Young, J. L. Jenkins, M. Glick, R. C. Glen and A. Bender, *J. Proteomics*, 2011, **74**, 2554–2574.
3 S. Whitebread, J. Hamon, D. Bojanic and L. Urban, *Drug Discovery Today*, 2005, **10**, 1421–1433.
4 R. B. Rothman, M. H. Baumann, J. E. Savage, L. Rauser, a. McBride, S. J. Hufeisen and B. L. Roth, *Circulation*, 2000, **102**, 2836–2841.
5 E. Lounkine, M. J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, E. Weber, A. K. Doak, S. Côté, B. K. Shoichet and L. Urban, *Nature*, 2012, **486**, 361–367.
6 M. S. Saporito, C. A. Lipinski and A. G. Reaume, in *Drug Repositioning: Bringing New Life to Shelved Assets and Existing Drugs*, ed. M. J. Barratt and D. E. Frail, John Wiley & Sons, Inc., Hoboken, New Jersey, 2012, part II, ch. 9, p. 279.
7 S. Ekins, A. J. Williams, M. D. Krasowski and J. S. Freundlich, *Drug Discovery Today*, 2011, **16**, 298–310.
8 T. T. Ashburn and K. B. Thor, *Nat. Rev. Drug Discovery*, 2004, **3**, 673–683.
9 M. Kuhn, M. Campillos, P. González, L. J. Jensen and P. Bork, *FEBS Lett.*, 2008, **582**, 1283–1290.
10 A. Bender, J. Scheiber, M. Glick, J. W. Davies, K. Azzaoui, J. Hamon, L. Urban, S. Whitebread and J. L. Jenkins, *ChemMedChem*, 2007, **2**, 861–873.
11 M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin and B. K. Shoichet, *Nat. Biotechnol.*, 2007, **25**, 197–206.
12 M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet and B. L. Roth, *Nature*, 2009, **462**, 175–181.
13 Nidhi, M. Glick, J. W. Davies and J. L. Jenkins, *J. Chem. Inf. Model.*, 2006, **46**, 1124–1133.
14 Y. Z. Chen and D. G. Zhi, *Proteins*, 2001, **43**, 217–226.
15 E. Kellenberger, N. Foata and D. Rognan, *J. Chem. Inf. Model.*, 2008, **48**, 1014–1025.
16 Y. Y. Li, J. An and S. J. M. Jones, *PLoS Comput. Biol.*, 2011, **7**, e1002139.
17 S. Zhao and S. Li, *PloS One*, 2010, **5**, e11764.
18 J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. a. Armstrong, S. J. Haggarty, P. a. Clemons, R. Wei, S. a. Carr, E. S. Lander and T. R. Golub, *Science*, 2006, **313**, 1929–1935.
19 J. Lamb, *Nat. Rev. Cancer*, 2007, **7**, 54–60.
20 M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen and P. Bork, *Science*, 2008, **321**, 263–266.
21 F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi and D. di Bernardo, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 14621–14626.
22 F. Iorio, R. Tagliaferri and D. di Bernardo, *J. Comput. Biol.*, 2009, **16**, 241–251.
23 G. Hu and P. Agarwal, *PloS One*, 2009, **4**, e6536.
24 M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. a. Morgan, A. Sweet-Cordero, J. Sage and A. J. Butte, *Sci. Transl. Med.*, 2011, **3**, 96ra77.
25 J. T. Dudley, M. Sirota, M. Shenoy, R. Pai and S. Roedder, *Sci. Transl. Med.*, 2011, **3**, 96ra76.
26 E. J. Cosgrove, Y. Zhou, T. S. Gardner and E. D. Kolaczyk, *Bioinformatics*, 2008, **24**, 2482–2490.
27 E. Jacoby, *Mol. BioSyst.*, 2006, **2**, 218–220.
28 C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo and D. S. Wishart, *Nucleic Acids Res.*, 2011, **39**, D1035–D1041.
29 M. Whirl-Carrillo, E. M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, C. F. Thorn, R. B. Altman and T. E. Klein, *Clin. Pharmacol. Ther. Ser.*, 2012, **92**, 414–417.
30 J. von Eichborn, M. S. Murgueitio, M. Dunkel, S. Koerner, P. E. Bourne and R. Preissner, *Nucleic Acids Res.*, 2011, **39**, D1060–D1066.
31 M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen and P. Bork, *Mol. Syst. Biol.*, 2010, **6**, 343.
32 D. Nitsch, J. P. Gonçalves, F. Ojeda, B. de Moor and Y. Moreau, *BMC Bioinf.*, 2010, **11**, 460.
33 D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen and C. von Mering, *Nucleic Acids Res.*, 2011, **39**, D561–D568.
34 J. Amberger, C. Bocchini and A. Hamosh, *Hum. Mutat.*, 2011, **32**, 564–567.
35 Z. Lu, *Database*, 2011, **2011**, baq036.
36 A. Ceol, A. Chatr Aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli and G. Cesareni, *Nucleic Acids Res.*, 2010, **38**, D532–D539.
37 T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan,

 This journal is © The Royal Society of Chemistry 2013

A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady and A. Pandey, *Nucleic Acids Res.*, 2009, **37**, D767–D772.

38 C. Alfarano, C. E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo, M. J. Dumontier, M. R. Dumontier, V. Earles, R. Farrall, H. Feldman, E. Garderman, Y. Gong, R. Gonzaga, V. Grytsan, E. Gryz, V. Gu, E. Haldorsen, a. Halupa, R. Haw, a. Hrvojic, L. Hurrell, R. Isserlin, F. Jack, F. Juma, a. Khan, T. Kon, S. Konopinsky, V. Le, E. Lee, S. Ling, M. Magidin, J. Moniakis, J. Montojo, S. Moore, B. Muskat, I. Ng, J. P. Paraiso, B. Parker, G. Pintilie, R. Pirone, J. J. Salama, S. Sgro, T. Shan, Y. Shu, J. Siew, D. Skinner, K. Snyder, R. Stasiuk, D. Strumpf, B. Tuekam, S. Tao, Z. Wang, M. White, R. Willis, C. Wolting, S. Wong, a. Wrong, C. Xin, R. Yao, B. Yates, S. Zhang, K. Zheng, T. Pawson, B. F. F. Ouellette and C. W. V. Hogue, *Nucleic Acids Res.*, 2005, **33**, D418–D424.

39 L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie and D. Eisenberg, *Nucleic Acids Res.*, 2004, **32**, D449–D451.

40 C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski and M. Tyers, *Nucleic Acids Res.*, 2011, **39**, D698–D704.

41 S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard and H. Hermjakob, *Nucleic Acids Res.*, 2012, **40**, D841–D846.

42 M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu and Y. Yamanishi, *Nucleic Acids Res.*, 2008, **36**, D480–D484.

43 D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio and L. Stein, *Nucleic Acids Res.*, 2011, **39**, D691–D697.

44 C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay and K. H. Buetow, *Nucleic Acids Res.*, 2009, **37**, D674–D679.

45 M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, The Gene Ontology Consortium, *Nat. Genet.*, 2000, **25**, 25–29.

46 T. Rattei, P. Tischler, S. Götz, M.-A. Jehl, J. Hoser, R. Arnold, A. Conesa and H.-W. Mewes, *Nucleic Acids Res.*, 2010, **38**, D223–D226.

47 I. Kondor and J. Lafferty, Proceedings of the nineteenth international conference on machine learning, Sydney, 2002, 315–322.

48 H. Yang, I. King and M. R. Lyu, Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, Amsterdam, 2007, 431–438.

49 E. Nabieva, K. Jim, A. Agarwal, B. Chazelle and M. Singh, *Bioinformatics*, 2005, **21**(Suppl. 1), i302–i310.

50 T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. a. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter and R. Edgar, *Nucleic Acids Res.*, 2009, **37**, D885–D890.

51 L. Gautier, L. Cope, B. M. Bolstad and R. a. Irizarry, *Bioinformatics*, 2004, **20**, 307–315.

52 B. S. Carvalho and R. a. Irizarry, *Bioinformatics*, 2010, **26**, 2363–2367.

53 G. K. Smyth, in *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, ed. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry and W. Huber, Springer, New York, 1st edn, 2005, part V, ch. 23, pp. 397–420.

54 Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, Z. Zhou, L. Han, K. Karapetyan, S. Dracheva, B. a. Shoemaker, E. Bolton, A. Gindulyte and S. H. Bryant, *Nucleic Acids Res.*, 2012, **40**, D400–D412.

55 A. Gaulton, L. J. Bellis, a. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2012, **40**, D1100–D1107.

56 T. Liu, Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson, *Nucleic Acids Res.*, 2007, **35**, D198–D201.

57 J. A. Hanley and B. J. McNeil, *Radiology*, 1982, **143**, 29–36.

58 M. Iskar, M. Campillos, M. Kuhn, L. J. Jensen, V. van Noort and P. Bork, *PLoS Comput. Biol.*, 2010, **6**, e1000925.

59 H. N. Chua, W.-K. Sung and L. Wong, *Bioinformatics*, 2006, **22**, 1623–1630.